OSU Working Papers in Linguistics 45 (11-35)

Vocal tract evolution and vowel production*

K. Bretonnel Cohen
kcohen@ling.ohio-state.edu

Abstract: Evolution has left the anatomically modern human with a supralaryngeal airway which is qualitatively different from that of all other animals. The source-filter theory of speech (the dominant theory of speech production in modern phonetics) relates articulations to their acoustic outputs by means of the action of a "filter"—the supralaryngeal vocal tract—on a "source"—air coming from a vibrating larynx. Since the human "filter" is qualitatively different from that of other animals, we should expect that the acoustic outputs that a human could generate would be different from those that other animals can produce. However, no satisfactory account of the nature of these differences has yet been given. A computer model was used to calculate variations in vocal tract transfer functions in human and non-human supralaryngeal airways as the location of a vocalic constriction is varied. The results suggest that the nature of the acoustic differences between the human and non-human anatomies has to do with nonlinearities which characterize articulatory/acoustic relations for the anatomically modern human vocal tract more so than for the non-human supralaryngeal airway. These differences arise from the range of articulations which are producible in a vocal tract with a bend in one wall, as opposed to one where both walls are straight. For the high vowels /i/ and /u/, these nonlinearities lead to areas of formant stability in the human airway which are significantly larger than those in the non-human airway. For the low vowel, the results suggest that the non-human airway should not be able to produce a front/back contrast between low vowels. In contrast, modelling of the modern human vocal tract correctly predicts the possibility of a front/back contrast for low vowels, though without any increase in areas of formant stability. The extrinsic tongue musculature of the human was then compared with that of *Pan troglodytes*. Using a perturbation theory model to evaluate the acoustic effects of extrinsic tongue muscle activity, it was found that the ability to generate the vowel triangle is related to the functional potentials of the extrinsic tongue musculature. These are not acoustically significant in the non-human. In the human, they give rise to the ability to generate the extreme points of the vowel triangle.

---

* Mary Beckman provided ideas, support, criticism, encouragement, and more. Cathy Callaghan introduced me to the evolution of language. Ashok K. Krishnamurthy derived the formulae for the modelling study, and Tzyy-Ping Jung and Michael J. Collins programmed them in Matlab. Stan Ahalt and the Center for Cognitive Science provided financial support. John A. Negulesco of the College of Medicine reviewed an earlier version of the section on the evolution of the vowel triangle and examined many bisected heads with me. Leslie Kent provided valuable discussion at all hours of the day and night. Frederick Parkinson listened to the whole thing many times, and many others provided valuable comments at the Spring 1993 meeting of the Acoustical Society of America, the 1993 Linguistic Institute, and the 1994 meeting of the Linguistic Society of America.

I. Introduction

I.1 The issue

The source-filter theory of speech relates the articulations by which speech is produced to the acoustic outputs which they generate. This relationship is understandable as the effect of a "filter"—the supralaryngeal vocal tract—acting on a "source," which in the prototypical case consists of periodically disturbed air exiting from a phonating larynx.

The supralaryngeal airway of the adult human is not just quantitatively but qualitatively different from that of all other animals, including our archaic hominid ancestors. Since we have qualitatively different "filters," we should expect there to be qualitative differences in the nature of the acoustic outputs that we can generate. However, a satisfactory analysis of the nature of these differences has not previously been presented. I will show that they are related to (1) nonlinearities in the relationship between articulatory and acoustic parameters, and (2) the contrasts that can be generated by the two sorts of filters.

I.2 The anatomy

The human vocal tract differs from that of other animals as a result of two trends in hominid evolution. These trends consist of flexion of the base of the skull and a decrease in the height of the larynx.

The basicranium, or base of the skull, is formed from the occipital, temporal, sphenoid, vomer, palatine, and maxillary bones (Jacob and Francone 1965:81). Together they form the bottom of the cranial vault, the roof of the mouth, and the superior boundary of the vocal tract. They serve as articulators (e.g. the alveolar ridge of the maxillary bone) and as the superior (in the anatomical sense of that word, meaning located towards the head) insertions of a number of the muscles of the vocal tract, including muscles of the velum, uvula, and pharynx (McMinn and Hutchings 1977:16).

All non-hominid animals have a relatively flat basicranium. In other words, the plane which is oriented along the base of the skull is relatively flat. Archaic hominids, like other animals, have a somewhat less flat, but still flat, basicranium. Over the course of the evolution of hominids from Australopithecus to Homo sapiens, basicranial flexion—the degree to which the basicranium is non-flat—increases gradually. With the appearance of Homo sapiens, it becomes bent, or flexed. For clear illustrations of the relevant structures, see the illustrations in, e.g., Lieberman (1975), (1984), and (1991). The chimpanzee is representative of the standard non-human mammalian basicranial shape. Note that the basicranial line of the human contains a sharply acute angle, while that of the chimpanzee does not.

All animals other than anatomically modern humans, then, have an unflexed skull base. All animals other than humans also have a larynx located high in the throat. In most mammals, the superior edge of the larynx is roughly parallel to the first cervical vertebra. In humans, the superior edge of the larynx is parallel to the 4th cervical vertebra. In the absence of soft tissue remains, the same sort of fossil record that is present for the development of basicranial flexion over the course of

hominid evolution is not present for laryngeal descent. However, comparative zoological evidence demonstrates clearly that in all other animals extant today, laryngeal height is inversely correlated with the degree of basicranial flexion. That is, the lesser the degree of basicranial flexion, the higher is the larynx in the neck (Laitman 1984, Laitman and Reidenberg 1988). So, if our hominid ancestors did not have high larynges to go with their unflexed basicrania, then they differed from all other animals in this respect. There is no reason to think that this is the case; probably, then, they did have high larynges.

The combined effect of these changes has been to leave the modern human with a pharyngeal cavity, located at a right angle to the oral cavity, which is absent in other animals. This is illustrated schematically very nicely in Hoffman *et al.* (1989:106). Thus, the trends of basicranial flexion and laryngeal descent over the course of the evolution of the hominidae have left us with a supralaryngeal airway which is qualitatively different from that of all other animals.

I.3 Lieberman's "abrupt discontinuity" analysis

Philip Lieberman pioneered the study of the relationship between the evolution of the human vocal tract and the phylogeny of human language. He has suggested that the acoustic significance of basicranial flexion and laryngeal descent lies in the differing articulatory capabilities of an airway consisting of a tongue located within an unbent tube as compared to an airway consisting of a tongue located within a bent tube. Lieberman (1975:115, 1984:278-280) interprets the modelling studies in Stevens (1972:57) as demonstrating that (quantal) vowel production requires abrupt discontinuities in cross-sectional area. This effect can be achieved—in the modern human vocal tract—by displacing the body of the tongue anteriorly and superiorly, as in the production of the high front vowel [i]; posteriorly, as in the production of the low back vowel [ɑ]; or superiorly and posteriorly, as in the production of the high back vowel [u]. According to Lieberman, this sort of abrupt discontinuity in cross-sectional area cannot be achieved in a straight-tube, standard non-human mammalian airway; rather, only gradual discontinuities can be achieved, with the tongue sloping gradually into and out of a constriction.

However, evidence from a variety of sources suggests that Lieberman is wrong. It may be the case that non-human airways cannot generate abrupt discontinuities in cross-sectional area. However, it is clearly the case that in the production of vocalic sounds, human speakers do not produce such discontinuities, either. Consider Figure 1, which shows (a) the sort of constriction whose lack of abrupt discontinuities in cross-sectional area, Lieberman claims, prevents the non-human vocal tract from producing the three "point" vowels, and (b) a tracing from a sagittal x-ray of a human speaker producing the vowel [u]. X-ray (e.g. Fant 1960, Perkell 1969), MRI (e.g. Moore 1992), and palatographic and ultrasound studies (e.g. Stone et al. 1992) all clearly show exactly what Lieberman posits for other animals, to the exclusion of humans: a tongue surface sloping gently into and out of a constriction. In fact, a variety of writers have commented on the gradual nature of vocalic constrictions, e.g.:

> Real constrictions, formed by the tongue in the vocal tract during natural speech, have a gradual shape.
>
> Mrayati *et al.* 1988:270

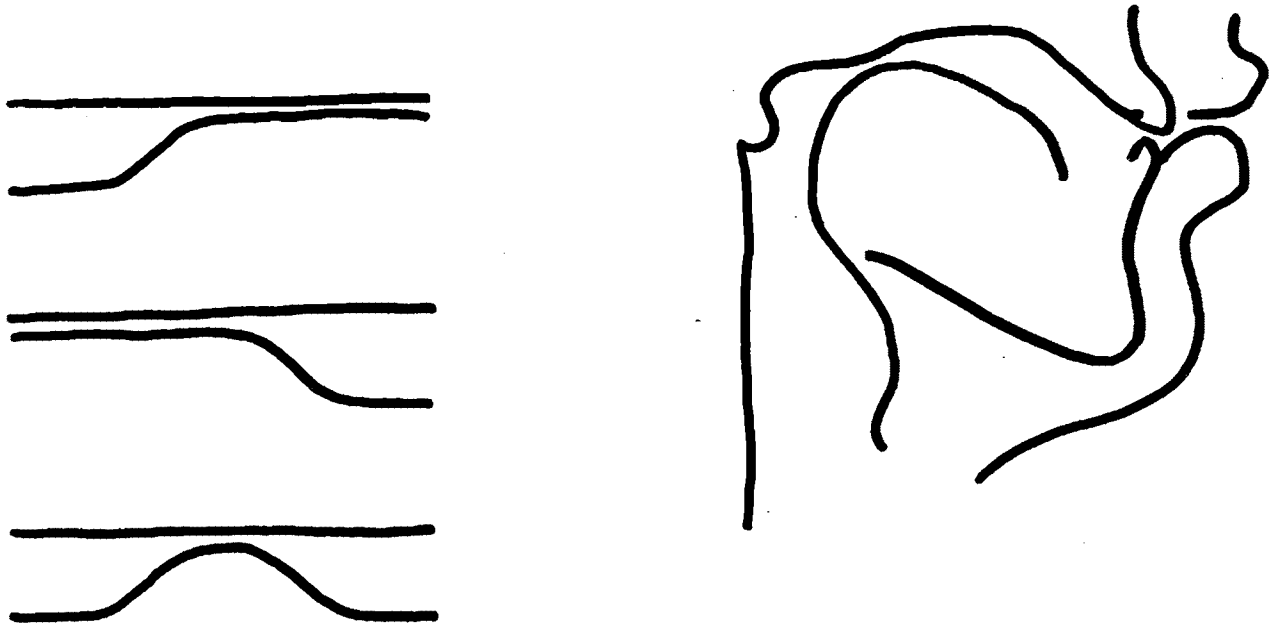

Figure 1. (a) Constrictions without abrupt discontinuities in cross-sectional area, adapted from Lieberman (1991).
(b) Tracing from a sagittal x-ray of a human speaker producing the vowel [u], adapted from Perkell (1969). The tongue, hard and soft palates, and posterior pharyngeal wall are shown.

> During the articulation of vowels, the tongue usually forms a constriction or region of minimum cross-sectional area....On either side of the constriction there is a gradual increase in area.
>
> Stevens and House 1955:485

Badin *et al.* talk at some length about this issue (1990:1297), though without providing quotable quotes. Furthermore, there is a history of modelling vocalic constrictions with gradual, nonabrupt constrictions which extends back through Fant's use of horn-shaped (i.e., tapered) tubes (1960:30 and following, 9 and following) through Stevens and House's use of a parabolic constriction (1955:486) as far as Chiba and Kajiyama (1941:82-83 and elsewhere).

So, whatever the acoustic consequences of the evolution of the human vocal tract may be, it seems clear that they cannot be related to the ability—or lack thereof—to produce abrupt discontinuities in cross-sectional area during vowel production. What, then, are they?

I.4 An "increased nonlinearity" analysis

Gunnar Fant's Acoustic theory of speech production (1960) explains the characteristics of the speech signal in terms of the output of a filter—the supralaryngeal vocal tract—acting upon a laryngeal source. The acoustic theory of speech production makes certain predictions about the way that the transfer function (i.e., the acoustic output, expressed as a set of formant frequencies, of a given vocal tract configuration) of the vocal tract should vary as a constriction is moved from location to location. These predictions are expressed in graphs called nomograms. Ladefoged and Bladon (1982) tested these predictions by producing sustained vocoids while varying only the place of constriction, using mirrors, bite blocks, and ultrasound to keep lip aperture, area of constriction, etc. constant. They noted that the formant structures produced by actual speakers varied from those predicted by Fant's nomograms. Specifically (among other things), at very forward (i.e., close to the lips, or far from the glottis) locations for an [i]-like constriction, the second formant frequency did not fall, as Fant's nomograms predicted. Rather, the second formant frequency stayed relatively stable.

Ladefoged and Bladon hypothesized that this effect was related to the fact that within the range of locations for a high front vowel constriction, the second formant is a back-cavity resonance. They suggest that "because of the curvature of the vocal tract, the length of this cavity does not increase when the constriction moves closer to the alveolar ridge" (p. 194). Rather, past the bend in the vocal tract, the back cavity length remains constant. While front cavity length decreases, there is a concomitant increase not in the length of the back cavity, but in its diameter. Though they were referring to the curve at the alveolar ridge, we were inspired by their comment to consider the effect of the curvature of the vocal tract as a whole, specifically the different effects of changing the location of a vocalic constriction in a tube bounded by a straight wall—analogous to the non-human supralaryngeal airway—versus the effect of changing the location of a vocalic constriction in a tube bounded on one side by a bent wall—analogous to the anatomically modern human supralaryngeal airway.

A geometric relationship such as that described by Ladefoged and Bladon can exist for a tube with a right-angle bend in it. However, it cannot exist for an unbent tube. Rather, in an unbent tube the only possible relationship between the lengths of the front and back cavities is that of a trade-off: as the back cavity length increases, the front cavity length decreases. And, in neither cavity is the diameter affected by variations in the length of the other.

Ladefoged and Bladon noted that "it is difficult to relate acoustic behaviors such as... formant discontinuities to the articulatory states which produced them, namely moving the tongue progressively in small steps along the upper surface of the vocal tract" (p. 192). This is certainly true, if our expectation is that articulatory/acoustic relationships should be linear. However, that is not necessarily the case. A non-linear relationship such as this is just the sort that would be predicted by Stevens' Quantal Theory of speech.

One of the hypotheses proposed here is that one of the acoustic consequences of the anatomical changes which occurred in the evolution of the modern human vocal tract can be described as a trend toward increasing the nonlinearity of the relationship between articulations and the associated acoustic output in the production of vowels. Stevens's (1972, 1989) modelling studies predict the existence of nonlinearities in articulatory/acoustic relations. Specifically, he claims that certain vowels are articulated in locations where there is a nonlinear relationship between the location of a constriction and the associated formant frequencies. For example, within the range of locations in which a high front vowel is produced, the second formant frequency is stable over a range of values for location. Stevens demonstrates these nonlinearities by means of a model of vowel production in which the location of a constriction is varied from the back to the front of the vocal tract by trading off the lengths of the front and back cavities.

This is precisely the relationship between front and back cavity lengths of which an unbent vocal tract is capable. It does in fact yield a nonlinear relationship between location of a constriction and second formant frequency over some range of values for constriction location. One wonders, then, if a bent-tube vocal tract has the same sort of nonlinear relationship between articulatory and acoustic parameters. If it turned out that there were no differences in the outputs generatable by straight-tube (non-human) and bent-tube (qualitatively different modern human) vocal tracts, that finding would be embarrassing for the source-filter theory. If it turned out that the bent-tube vocal tract is more linear than the straight-tube vocal tract in its articulatory/acoustic relationships, that would be an embarrassment for Stevens' Quantal Theory and would constitute an absolute refutation of the thesis of this paper. If, on the other hand, it turned out that the bent-tube vocal tract is associated with a less linear articulatory/acoustic relationship—e.g., if there were a larger area of stability for the second formant in the range of locations in which high front vowels are articulated—that would support the thesis that the human vocal tract has evolved in the direction of increased articulatory/acoustic nonlinearities.

II. Method

I tested this hypothesis by modelling vowel production in straight and bent vocal tracts. Transfer functions were calculated with a transmission line analogue model. The model was tested against the comparable nomograms in Stevens

(1989), and against actual production data as part of a separate study (Beckman *et al.* (1995).

I modelled vowel production in a straight-tube, non-human vocal tract as in Stevens (1989). The location of the constriction is varied by trading off front and back cavity length: i.e., as the length of the back cavity increases, the length of the front cavity decreases.

I modelled vowel production in a bent-tube, modern human vocal tract by varying the sizes of the front and back cavities in a way which more accurately reflects the geometry of the modern human vocal tract. I used the same ratios of cavity sizes and constriction characteristics as in Stevens 1989. This sacrifices some realism, but has the advantage of allowing direct comparison with Stevens's classic model. For [i] and [u], I varied back cavity length from the glottis to the point corresponding to the bend in the vocal tract by trading back and front cavity lengths. However, past the bend, I varied the constriction location by decreasing front cavity length while keeping back cavity length constant and increasing back cavity diameter. This is expressed schematically in Figure 2 (b) below. The motivation for this this approach is pictured in Figure 2 (a). If a tongue body shape like that pictured in Figure 2 is moved forward, the effect on the back cavity is an increase not in distance along the vertical axis—i.e., back cavity length—but in distance along the horizontal axis, i.e. back cavity diameter.

The novelty of my approach lies in its strategy for varying the location of the constriction. This strategy differs from that of other models in that it explicitly recognizes the different effects on cavity area of changes of location in the anterior portion of the vocal tract as compared to changes of location in the posterior portion of the vocal tract. For a review of other models, all of which fail to take these differing effects into account, see Section VI, *Postscript I: vocal tract models.*

II.1 Changes in overall length

My strategy for modelling changes in place of articulation in the modern human vocal tract results in a net decrease in the overall length of the vocal tract. This occurs as the constriction is advanced from the mid-point to the front of the vocal tract. This might make one wonder if the effect of second formant frequency stability is an artifact of the decrease in overall length. This can be shown not to be the case. Reduction of overall length with the same ratio of front and back tube lengths results in a pattern of increase in all frequencies. Our findings show an increase in the third formant frequency only, with the second formant remaining stable and the first formant essentially unaffected. This point will be repeated below after presentation of the modelling results.

The reduction in length in our model is in the amount of 3 centimeters. There is justification from physiological studies for some reduction in length in moving from a back to a front vowel articulation, if not in the same magnitude. Adjusting the modelling strategy so that the reduction in length more closely approximates a physiologically reasonable amount results in a pattern of second formant frequency behavior which is substantially similar to those observed in my model of the modern human vocal tract, the differences being in the range (in articulatory and acoustic space) of the stable region. Thus, the patterns of formant behavior generated by the model cannot be said to be due to inappropriate reduction of the vocal tract length. Furthermore, my model has the advantage of generating the actual patterns of formant behavior observed in human speakers (compare my

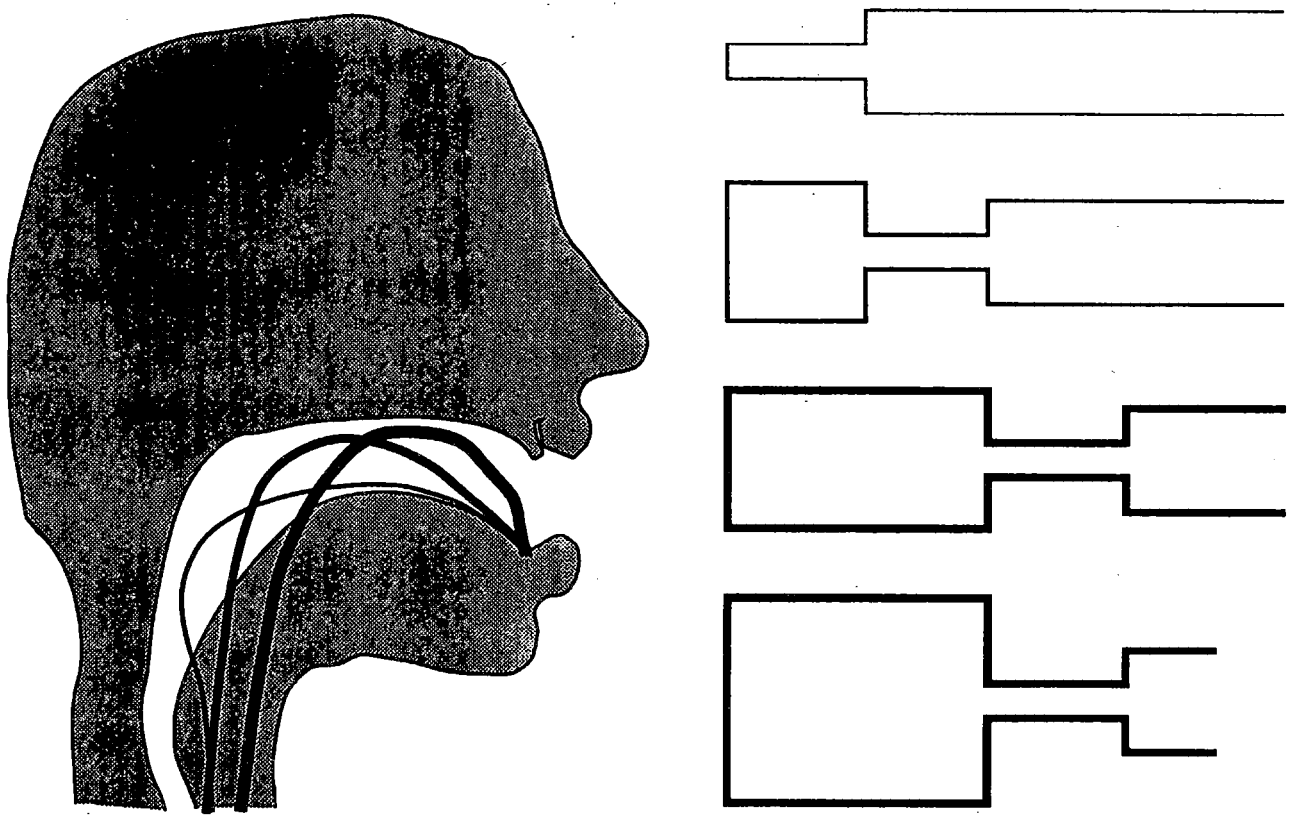

Figure 2. (a) Range of tongue shapes from back to front constriction.
(b) Schematic representation of variation in tube lengths and areas.

findings with, e.g., Ladefoged and Bladon) with fewer control parameters than an isolongitudinous model (one which maintains overall length) requires; it does a better job of generating them than do the isolongitudinous models, in the case of the third formant frequency. Kent *et al.* point out that "regardless of the individual approach taken, the basic goal in [modelling studies of articulatory/acoustic relations] has been to reduce the number of degrees of freedom" (1991:268). In that respect, the model I have used for the modern human vocal tract is superior to an isolongitudinous model.

III. Results

Figure 3 shows the effects of varying the location of an [i]-like constriction within a modern human vocal tract (solid line) versus a non-human supralaryngeal airway (broken line). The plot for the straight-tube, non-human configuration duplicates the relevant nomogram in Stevens (1989:12). The results for the bent-tube, modern human configuration differ in that this configuration yields a much larger area of stability for the second formant within the range of locations in which a high front vowel is articulated. With the non-human configuration, the second formant is even nominally stable only from perhaps 7.5 cm to 8.5 cm from the glottis. In contrast, in the modern human model, the second formant frequency varies hardly at all from 7.5 cm to 11 cm from the glottis.

For [u], I varied chamber lengths and widths as for [i]. A short tube was added to the anterior end for the lip-rounding component of [u], and the cavities anterior and posterior to the constriction were of equal width for locations in the posterior portion of the vocal tract.

The results for an [u]-like constriction are shown in Figure 4. Note that in the modern human model, over a range of values from around 8 cm to 10 cm from the glottis, the third formant remains high and distant from the second formant. In the non-human model, a backness percept can be expected only over the range of values for constriction location from about 4 cm to 8 cm from the glottis, where the second formant-third formant distance is large due to the second formant being low. At least half of this range is too far back to be at all realistic, and over most of it the first formant frequency is rather high for a high vowel. (The frequency of the first formant is inversely correlated with vowel height.) The modern human vocal tract allows for a backness percept over a wider range of values for constriction location, including more realistic values for location centered around 8 cm from the glottis and a range of locations where the first formant frequency is lower (as is appropriate for a high vowel). A large second formant-third formant distance is maintained not by keeping the second formant low but by keeping the third formant high even as the second formant begins to rise.

Modelling the production of [a] is an interesting challenge, and highlights the difficulty in describing low vowels within a constriction-based model of vowel production. [a]-like vowel configurations differ from non-low vowel configurations in that non-low vowels have a constriction and two cavities—one anterior and one posterior to the constriction. Constriction location is varied by changing the relative sizes of the two cavities; the characteristics of the constriction itself, i.e. its length and diameter, remain constant. In contrast, for a low vowel, there are not two cavities whose areas can be manipulated independently of the constriction. Rather, for a low back vowel, the "constriction" consists of the small

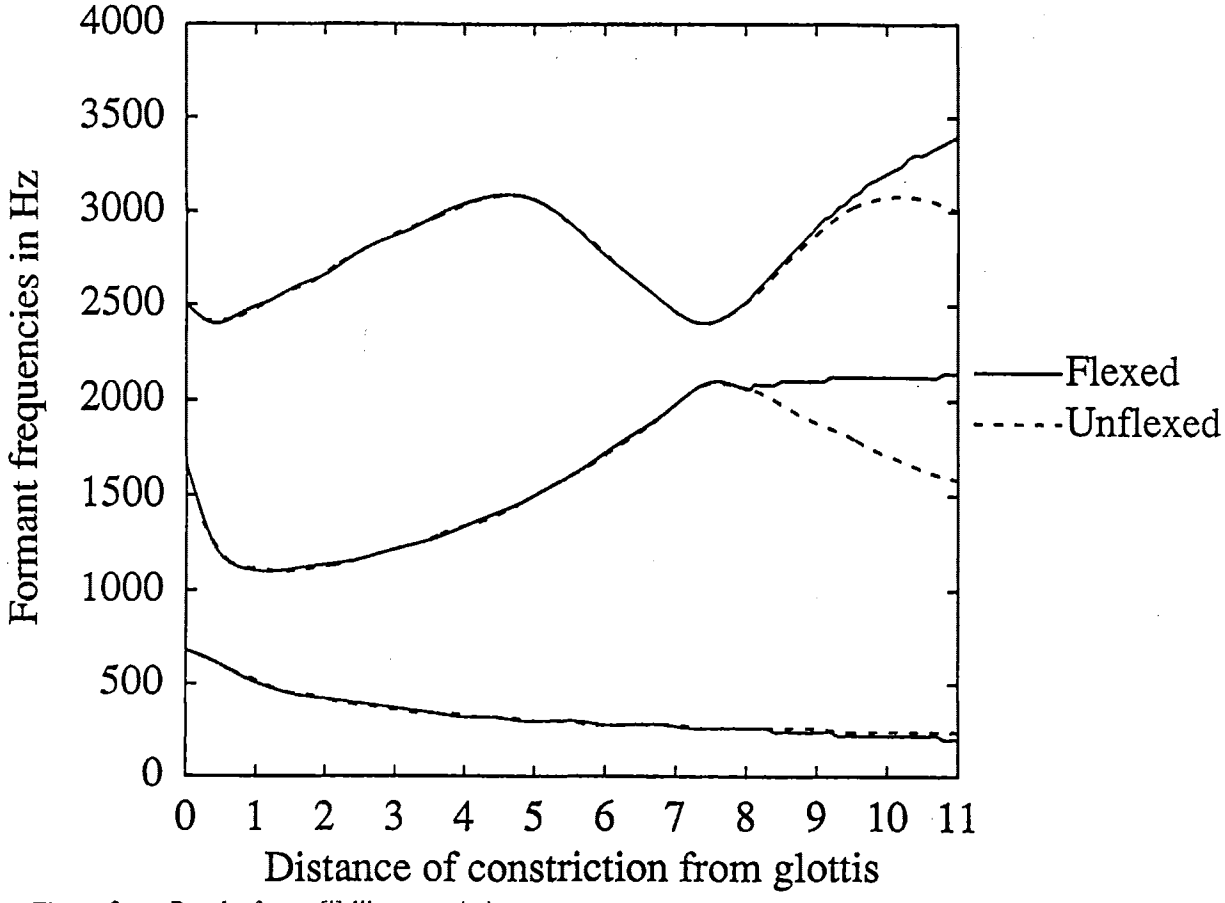

Figure 3. Results for an [i]-like constriction.

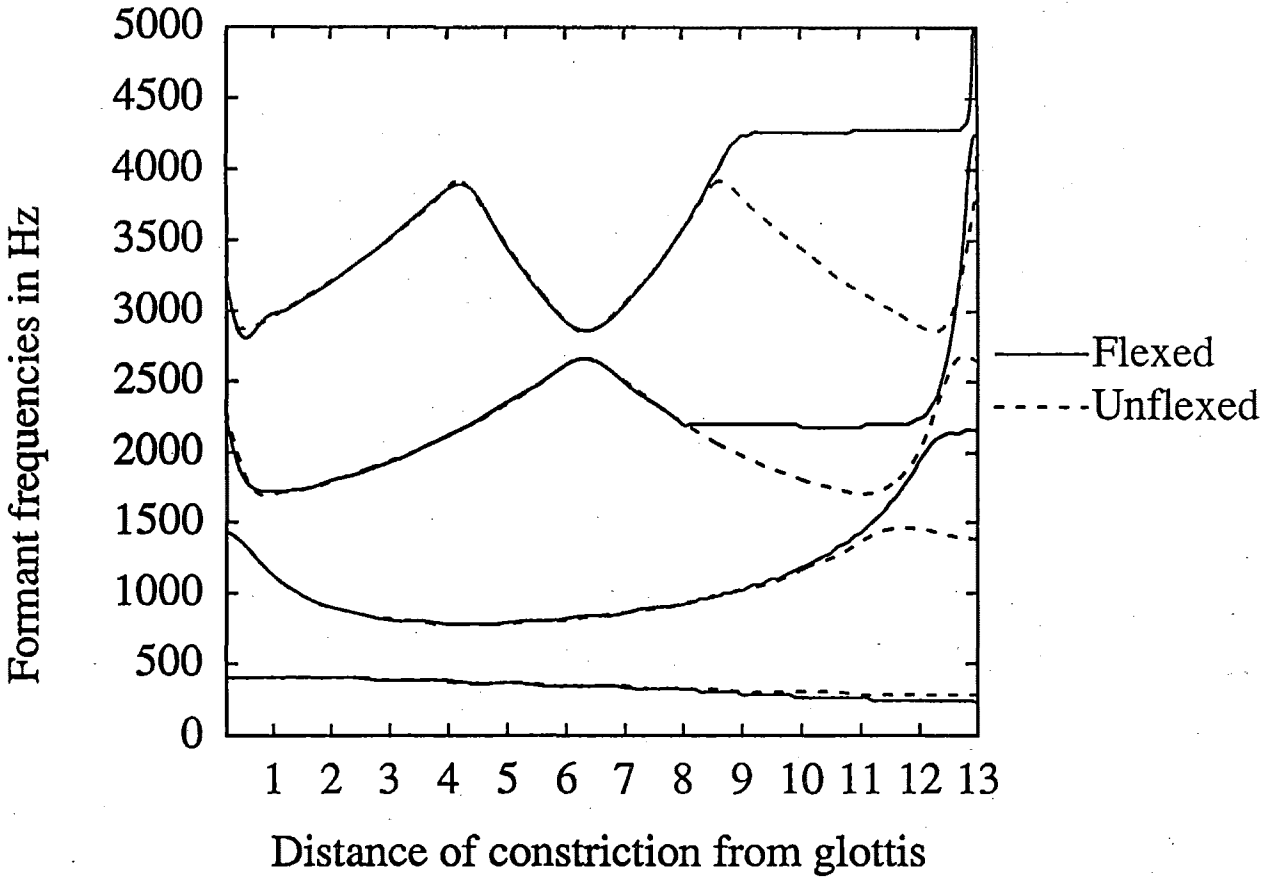

Figure 4. Results for an [u]-like constriction.

back cavity cross-sectional area. For a low front vowel, the vocal tract configuration approaches that of a neutral, i.e. completely non-constricted, tube.

Production of an [a]-like vowel was modelled with two cavities, with a large ratio of front:back cavity area. Constriction location was varied from the glottis to the bend by trading off front and back cavity lengths. Past the bend, back cavity diameter was increased as front cavity length was decreased. This was continued only until a neutral-tube-like configuration was achieved.

Figure 5 shows the results for an [a]-like constriction. Note that in the non-human model, the second formant frequency varies symmetrically with respect to the midpoint of the vocal tract. There is a single wide region wherein the first and second formants are close together and stable, the first formant being high and the second formant being low. This implies that there should only be a single low vowel, and that it should be a back vowel. In the modern human model, as the location of the constriction is moved forward past the bend, the second formant rises sharply. This allows for low vowels to contrast in front/backness, as in fact they do, e.g. English [ɑ] versus [æ].

III.1 Length reduction revisited

As mentioned above, one might wonder if the differences between the human and non-human modelling results are due simply to the overall reduction in length which occurs in the human model. This is clearly not the case. The effect of reduction of the overall length of the entire vocal tract would be a shifting upwards of all formant frequencies. In the human model, the reduction in length occurs as the constriction location is moved anteriorly from the midpoint of the vocal tract, so if overall length reduction were the cause of the differences between the human and non-human models, we would expect to observe identical patterns of formant change in both models as the constriction was moved from the glottis to the midpoint of the vocal tract, with all three formant frequencies increasing in the human model as the constriction was moved further forward from that point. Instead, we see a pattern explainable by (1) a length decrease affecting only the front tube, and (2) a concomitant increase in back cavity area. As the constriction is moved from the glottis to the midpoint, formant behavior is identical in both models. Past the midpoint, not all three formant frequencies, but rather only the third formant frequency, increase. The increase in the third formant frequency is the result of the shortening of the front tube, with which the third formant frequency is associated in this region of the vocal tract. The second formant frequency does not increase, but rather stays the same. This is the result of the lack of change in the length of the back tube, with which the second formant frequency is associated in this region of the vocal tract. The first formant frequency does not increase, but rather falls. This results from the fact that the first formant is a Helmholtz resonance (Johnson 1994), and as such is sensitive only to the characteristics of the constriction and the back cavity. Anterior to the midpoint, the constriction and back cavity length do not change, and therefore cannot be the cause of the drop in the first formant frequency as the constriction is moved forward from the midpoint: rather, the decrease in the first formant results from an increase in back cavity area, to which a Helmholtz resonance is sensitive. The increase in area in the human model is equivalent to the increase in length in the non-human model, so the behavior of the first formant frequency is the same in both cases.

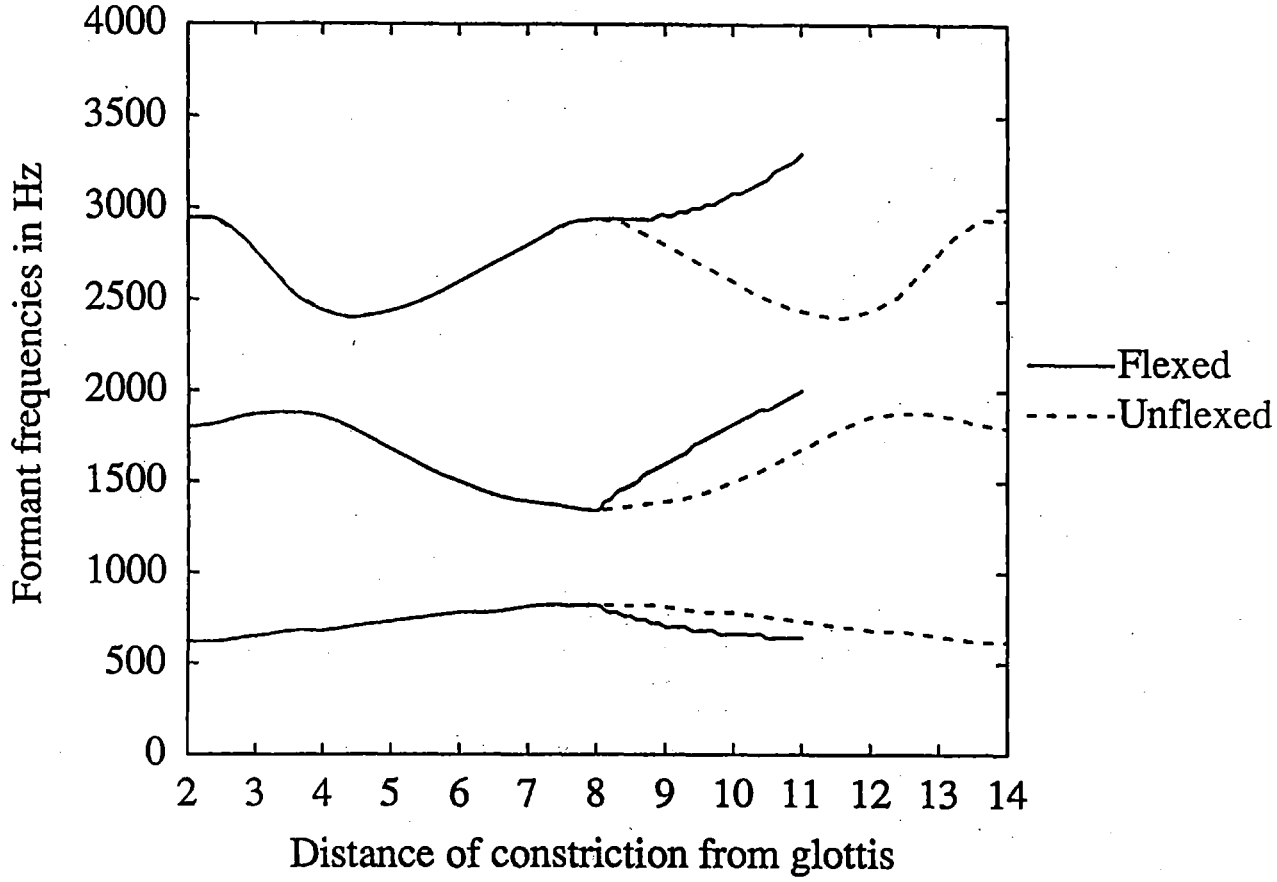

Figure 5. Results for an [a]-like constriction.

IV. The evolution of the vowel triangle

IV.1 Functional potentials, and perturbation theory

As the findings of the previous section show, acoustic stability in the production of vowels is one consequence of basicranial flexion and laryngeal descent. I will now show that another consequence of these changes has been the genesis of the ability to produce acoustic correlates of the distinctive features of vowel height and backness which are the materials of the systems of vowel contrasts widely attested in human languages. In non-human animals, the functional potentials (directions of movement which could be caused by contraction) of the extrinsic tongue muscles are in opposition to each other. However, in the non-human, activity of the extrinsic tongue musculature should not be expected to produce effects on relative formant frequency values. (They might change overall structure, e.g. producing an overall lowering of formant frequencies—but not a change in the frequency of one formant relative to the others.) In contrast, in anatomically modern humans, the oppositions between the functional potentials of the extrinsic tongue muscles, combined with the availability of a pharyngeal cavity into which the tongue may be displaced, allow for the production of the acoustic distinctive features of vowels by the activity of the extrinsic tongue musculature.

Work by Honda and his cohorts (e.g. Kakita *et al.* 1985, Honda *et al.* 1993, Honda 1994) suggests that contrasts in vowel sounds are made possible by oppositions in functional potential between the genioglossus, styloglossus, and hyoglossus muscles, and that this opposition is made possible by morphological changes in the supralaryngeal airway in the course of hominid evolution. However, besides their suggestions, this topic remains unexplored. In this paper, I compare the potential oppositions present in the human vocal tract with those which appear to be present in the non-human supralaryngeal airway. I use published data on chimpanzee anatomy from Swindler (1973) to demonstrate the differences in functional potential between the human and non-human anatomical conditions. I then use a perturbation model of the acoustics of vowel production to show the implications of these differences for speech. It will be seen that the orientation of the extrinsic tongue muscles in animals other than anatomically modern humans is preadaptive for speech. Oppositions in the orientations of these muscles relative to each other and relative to the tongue body as a whole have little or no acoustic consequences in a supralaryngeal airway which has an unflexed basicranium and a high larynx, and thus lacks a pharyngeal cavity. In contrast, once basicranial flexion and laryngeal descent occur, the oppositions of these muscles become acoustically consequential, enabling the production of, and oppositions between, the prototypically human vowels [i], [u], and [ɑ].

A comparison of the acoustic capabilities of the human and non-human vocal tracts can be made by comparing the sorts of deviations from a uniform tube that they can effect. The acoustic affects of these (varying) deviations from a uniform tube can then be compared by means of a perturbation-theory-based model of vocal tract acoustics. This sort of model owes much to the theory of vowel production developed in Chiba and Kajiyama (1941); I will use here a less familiar but more sophisticated model based on the work of Mrayati *et al.* (1988).

Mrayati *et al.*'s perturbation theory model is based on consideration of the effect of small, localized changes of cross-sectional area in the human vocal tract. Many other theoretical models use a transmission line analogue to model the vocal tract as a series of two tubes, with the length of each tube determining particular individual formant frequencies. In contrast, Mrayati *et al.*'s model works by consideration of the effects (on the resonance of the vocal tract as a whole) of changes in cross-sectional area on the potential and kinetic energy of air flowing through localized sections of the vocal tract. Following the work of Fant and Pauli, Mrayati *et al.* use these parameters to calculate the direction and amount of change in a formant's frequency from the area and the total, kinetic, and potential energies of a region (or subsection) of the vocal tract.[1] This approach to the acoustics of speech nicely models the formant transitions in CV (consonant-vowel) sequences, is fruitful in the consideration of the classic acoustic-articulatory inversion problem, and is a nice addition to the work of Stevens, in which the effect of changes in area, as opposed to changes in location, of a constriction receives rather short shrift. It also provides a nice account of the compensatory effects of changes in area at the lips and glottis. (I have some problems with the required orthogonality of changes in area elsewhere in the vocal tract. Such changes are clearly shown by Mrayati *et al.* to be acoustically orthogonal, but whether they can be articulatorily orthogonal in areas of the vocal tract other than its ends is not so true, I don't think.)

One finding of their work is that under the conditions which characterize (central) vowel production, i.e. relatively large "constriction" area and with the presence of acoustic coupling between the ante- and post-constriction tubes, there exist four regions of the vocal tract which have unique effects on formant structure, such that a constriction in one of these regions will produce a characteristic effect on each of the first two resonant frequencies of the vocal tract (i.e., formants). (The number of these "distinctive regions," as Mrayati *et al.* call them, is actually related to the number of formants being considered, so that for three formants, there are eight regions, for four formants, there are fourteen, etc.) These regions are illustrated schematically in Figure 6 (adapted from their Figure 2) with respect to a schematized tube. (Note that Figure 6 shows three formants, and therefore eight regions are identified on it.) (Mrayati *et al.*'s nomograms show the effect of *increasing* the area of a region, rather than the effects of constricting an area, as we are accustomed to seeing. I have adapted them to show the effects of a constriction within a region by inverting the original Figure 2.) Note that there are regions which produce a low F1 and high F2 (labelled C), low F1 and low F2 (labelled A), and high F1 and low F2 (labelled C-bar). These correspond to the vocal tract regions constricted in the production of the vowels [i], [u], and [a].

IV.2 The extrinsic tongue musculature

In both *Homo sapiens* and the chimpanzee, *Pan troglodytes,* the extrinsic tongue muscles consist of the genioglossus muscle, the styloglossus muscles, the hyoglossus muscles, and the palatoglossus muscles. (Negulesco 1993:63). All of the extrinsic muscles originate externally to the tongue, but have insertions within it. (In this they differ from the intrinsic tongue muscles, which have both their origins and their insertions within the tongue.) The palatoglossus muscle has not

[1] Section VII, *Postscript II: Mrayati et. al's perturbation theory* gives a detailed explanation of the model.

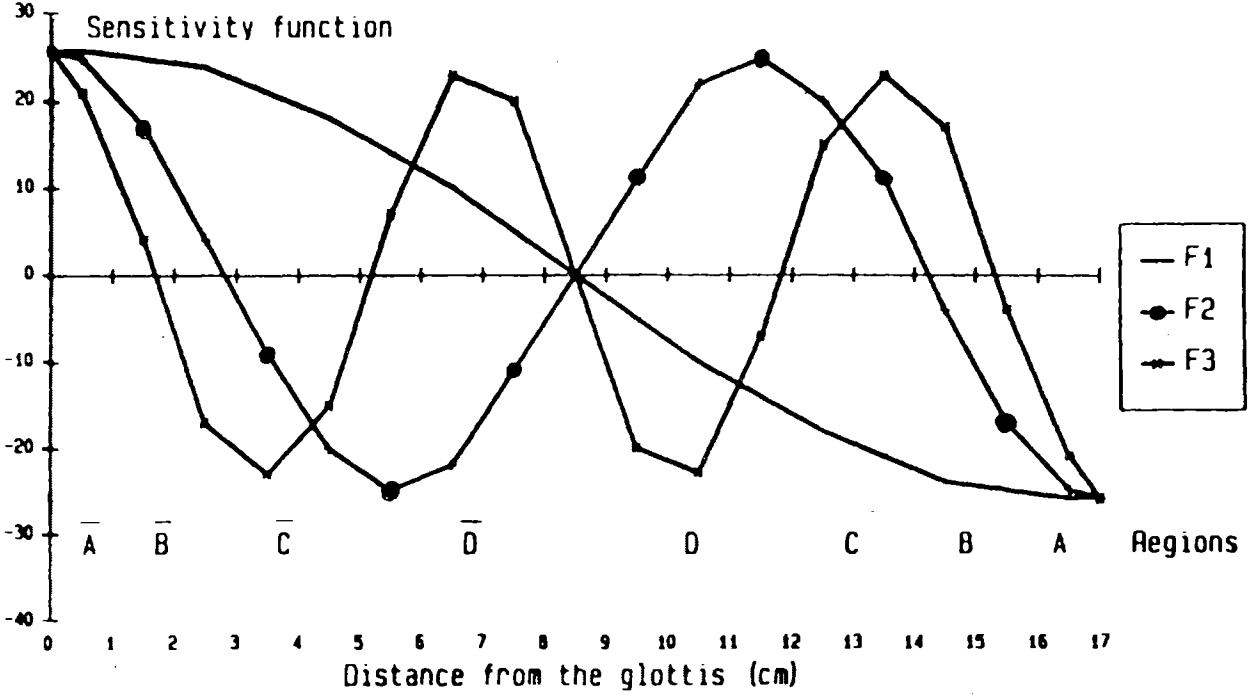

Figure 6. Distinctive regions in a perturbation model of vocal tract acoustics, adapted from Mrayati *et al.* (1988).

been shown to be of importance in the production of vowel sounds, and will not be discussed further here.

IV.2.a Genioglossus

In both the human and Pan, the fibers of the posterior portion of the genioglossus run roughly anterior-posterior. The genioglossus originates on the upper mental spines of the mandible and appears to form the main body of the tongue, when viewed in sagittal section. (I should note here that although the functional division of the genioglossus in the human is well-established on the basis of electromyographic data, such a division has not, to my knowledge, been demonstrated for Pan.) In modern Homo, the root of the tongue forms the anterior border of the pharyngeal cavity. In contrast, in Pan, the tongue is located completely within, and in fact nearly fills, the oral cavity.

The function of the genioglossus muscle in vowel production is to produce the constriction for high front vowels. Radiographic (Fant 1960, Perkell 1969) and MRI (Moore 1992) studies have demonstrated that the production of [i] is characterized by displacement of the tongue body superiorly and anteriorly in the oral cavity, so that the tongue dorsum is positioned beneath the hard palate, posterior to the alveolar ridge. Electromyographic studies (e.g. Smith 1971) of tongue muscle activity during speech production have demonstrated that this is accomplished by contraction of the posterior portion of the genioglossus muscle.

The location of the constriction corresponds to region C in Mrayati *et al.*'s model. The acoustic effect of a constriction at this point is a lowering of the first formant frequency and an elevation of the second formant frequency. This produces the formant pattern associated with the vowel [i].

In the chimpanzee, the orientation of the fibers of the posterior genioglossus suggests that the functional potential of the genioglossus m. would probably be to displace the tongue body anteriorly and superiorly, if the mandible were lowered to allow it room to move within the oral cavity. However, it is not at all clear that the chimpanzee tongue could be displaced forward, as it seems to be attached at its "root," rather than having a root which is a free-moving structure, as in modern Homo. Thus, genioglossus contraction has no obvious acoustic effect in the chimpanzee.

IV.2.b Hyoglossus

The hyoglossus originates from the greater horn of the hyoid and inserts into the body of the tongue. It appears to have the functional potential to move the tongue body posteriorly and inferiorly in both species. However, while Homo has space into which to move the tongue body in this direction—i.e., the pharyngeal cavity whose presence is the result of basicranial flexion and laryngeal descent—Pan does not.

The function of the hyoglossus muscles in vowel production is to produce the constriction for low (back?) vowels. The production of [ɑ] is characterized by displacement of the body of the tongue inferiorly and posteriorly. This is accomplished by contraction of the hyoglossus mm. (and possibly the pharyngeal constrictors). The location of the constriction corresponds to region C-bar in

Mrayati *et al.*'s model. The acoustic effect of a constriction at this point is a rasing of the first formant and lowering of the second formant. This produces the formant pattern associated with the vowel [ɑ].

In the chimpanzee, the hyoglossus muscles appear to have the functional potential to move the tongue body posteriorly and inferiorly. However, in the absence of a pharyngeal cavity, there appears not to be a space into which to move to the tongue body. Thus, hyoglossus contraction has no obvious effect in the chimpanzee.

IV.2.c Styloglossus

The styloglossus muscle is present in both species, originating on the styloid process of the temporal bone and inserting into the tongue body. However, its orientation relative to the genioglossus m. and the tongue body as a whole differs between the two species. In modern Homo, the styloglossus is oriented at an angle to the genioglossus, such that its functional potential is to move the tongue body posteriorly and superiorly. Styloglossus contraction causes the tongue body to move perpendicular to the long axis of the tube, producing a constriction around the area of the velum.

Radiographic studies demonstrate that the production of [u] is characterized by displacement of the tongue body superiorly and posteriorly. Electromyographic studies (Smith 1971:65) demonstrate that this is accomplished by contraction of the styloglossus mm. The location of the constriction corresponds to region D-bar in Mrayati *et al.*'s model. A vocalic constriction at this point has the effect of lowering the second formant frequency.

The location of the constriction is near a pressure minimum for both the first and second formant. The model therefore suggests that [u] should have low first and second formants, as it does. In the human, styloglossus muscle contraction produces a sound with a high first formant and low second formant by moving the tongue body perpendicular to the vocal tube, near its mid-point. In contrast, in Pan, styloglossus contraction does not move the tongue body perpendicular to the axis of the vocal tract, but rather parallel to it. If such movement is in fact possible in the absence of a cavity posterior to the tongue and oral cavity, the effect would be expected to be that of an overall lowering of formant frequencies. However, since the styloglossus muscles do not run perpendicular to any region of the chimpanzee airway, styloglossus contraction would not have the effect of changing the cross-sectional area of any region of the vocal tract, and thus it would not be expected to have an effect on any one formant frequency.

IV.3 Discussion

Vowels in human languages contrast in terms of two distinctive features: vowel height, and vowel backness. These distinctive features have acoustic correlates: the first formant frequency, for vowel height, and the second formant frequency, for vowel backness.

Distinctive features may be thought of as having some linguistic value by virtue of the presence of an opposition between their opposite values. The evidence reviewed above shows that production of these acoustic correlates of the distinctive

features for vowels can be related to the activity of the extrinsic tongue musculature. In the human, the opposition between the functional potentials of the genioglossus muscle and the hyoglossus muscles is related to the opposition between high and low first formant values, i.e. the acoustic correlates of low vowel height and high vowel height. The opposition between the genioglossus and the styloglossus is related to the opposition between high and low second formant values, i.e. the opposition between vowel frontness and vowel backness. These acoustic oppositions are generatable in the human, but not in the non-human, because the anatomical oppositions to which they are relatable are present in the human, but not in the non-human.

It will be noted that the effect of styloglossus contraction, causing a constriction in region D-bar, on the first formant is not to decrease it, as must be the case for a high vowel. The model portrays a connection between height and backness, such that front vowels are predicted to be high, and back vowels are predicted to be low. This fact makes it difficult to derive /u/ from extrinsic tongue muscle activity alone, but should not be seen as a weakness for Mrayati *et al.*'s model. Rather, the model suggests a reason why lip rounding, which *does* cause first formant lowering, is involved in the production of high back vowels, and indeed is redundant to [+back], in the vast majority of the world's languages. Consulting tables of formant values in a variety of sources (e.g. Ladefoged 1993:197 for American English; Shalev, Ladefoged, and Bhaskararao 1993:91 for Toda; Blankenship, Ladefoged, Bhaskararao, and Chase 1993:129 for Khonoma Angami; Bradlow 1993:24 for English and Spanish) it will be seen that high back vowels generally have lower F1's than do front vowels of the same height. This cross-linguistic pattern is not surprising, given the association between front vowel articulations and low first formant values which this model suggests.

The comparative anatomical evidence from Pan troglodytes suggests that potential opposition in functional potential of the extrinsic tongue musculature existed already in the archaic hominids: in Pan, the genioglossus could potentially move the tongue body in a posterior-anterior direction, while the styloglossus could potentially move it in an anterior-posterior direction. Similarly, the hyoglossus could potentially move the tongue body inferiorly, while the genioglossus and styloglossus could move it superiorly. Basicranial flexion changed the origins, or bony attachment points, of these muscles, and thus their functional potentials relative to each other. Laryngeal descent and the resultant presence of a pharyngeal cavity posterior and inferior to the oral cavity allowed for movement of the tongue in dimensions not previously possible—posteriorly in response to styloglossus contraction, inferiorly in response to hyoglossus contraction. The effect of basicranial flexion and laryngeal descent, then, is to allow these preadapted oppositions to have acoustic consequences, where before they did not.

V. Conclusion

The production of vocal language requires the ability to generate contrasts and the ability to produce contrasting sounds with stability. Our data show that the abilities to produce distinctions in vowel height and vowel backness are direct results of the evolutionary changes which have lead to the anatomically modern human having a supralaryngeal airway which is qualitatively different from that of all other animals. Perhaps more importantly, they suggest that the relationship between articulations and acoustic outputs in the modern human is characterized by regions of acoustic stability. These regions make it more possible to produce

speech: if there is a larger range of values for some articulatory parameter (constriction location, in these cases) which will yield some desired acoustic output, then one can be less precise in attempting to produce it. Note that the differences between the acoustic outputs of the two sorts of supralaryngeal airways arise not from the ability, or lack thereof, to produce any one particular sort of articulation, or any one particular vowel. Rather, they arise from the *ranges* of articulations producible in the two sorts of airways. Nor is it being claimed that the differences in acoustic outputs results from any *acoustic* effect of the bend in the airway: rather, they arise from the different range of *articulations* which result from the presence of a bend in one of the boundary walls of the airway.

It should be noted that although I disagree with Lieberman's interpretation of the acoustic significance of human vocal tract evolution, my findings are not incompatible with the theory of human language evolution which he has developed, and about which I am agnostic. His theory requires that there be some qualitative difference between human and non-human supralaryngeal airways, and substitution of the contents of this paper for the analogous section in any of his publications would harm his overall theory not a whit.

Nonlinearities in articulatory-acoustic relations have been the topic of much discussion in past years. Some have felt them not to be relevant to the description of speech, e.g. Ladefoged and Lindau (1989). Others have looked for explanations from some more general system, independent of the specific characteristics of speech, as do Abry *et al.* (1989). They are the expected finding within the framework of a theory that looks to nonlinearities in articulatory/acoustic relations to explain the workings of the phonetic/phonological component of the grammar. The work of Stevens (1972, 1989), Wood (1982, 1986), Beckman *et al.* (1995), and others suggests that human speakers utilize these sorts of articulatory/acoustic relations. My research suggests that they characterize articulatory/acoustic relationships in human, but not in non-human, airways.

VI. Postscript I: vocal tract models

Stevens and House (1955) varied the location of a constriction by changing the distance from the glottis of the high point of a parabola. Models such as those of Maeda (1990) and Lindblom and Sundberg (1971) are based on measurements of the dorsal surface of the tongue relative to the palate and posterior pharyngeal wall during the production of vowels. Constriction location is varied by extrapolating the points through which the tongue surface would pass when moving between the actual measured locations. Stevens (1972) and (1989) are based on a model in which the location of a constriction is varied by trading off the lengths of the cavities anterior and posterior to the constriction. Carré and Mrayati (1990) vary constriction location by means of successive transversal changes of cross-sectional area in a series of concatenated tubes. All such models have in common the characteristic of treating the variation of the location of a constriction in any one part of the vocal tract as a task just like variation of the location of a constriction in any other part of the vocal tract. They fail to capture the generalization that varying the location of a constriction affects the relative sizes of the cavities anterior and posterior to the constriction differently in the front and back regions of the vocal tract. Jackson (1988) provides a comprehensive review of models of vowel production, none of which take into account this aspect of the geometric relationships between the front and back regions of the vocal tract.

VII. Postscript II: Mrayati *et al.*'s perturbation theory

Mrayati *et al.*'s perturbation theory, and the theory of distinctive regions and modes which is derived from it, is based on consideration of the effects of cross-sectional area on the total, potential, and kinetic energy of small, local areas of the vocal tract. They use these parameters to calculate the effect on the resonances of the vocal tract as a whole of small, local changes in area. In these calculations, the total energy of the vocal tract is considered to be a constant, as is reasonable for the lossless case, i.e. if the loss of energy through damping by the soft tissue of the vocal tract, radiation out of the mouth, etc., is ignored.

In this theory, potential and kinetic energy are analogous to the familiar measures of a flowing gas: pressure and velocity. The vocal tract will resonate at frequencies related to the pressure waves whose wavelength is optimum for a vocal tract of a given length. These waves of different frequencies have pressure and velocity maxima and minima at different points along the length of the vocal tract. The familiar figures in Chiba and Kajiyama (1941) show the locations of the velocity maxima (marked N_n) and minima (at the points where lines cross). Pressure maxima are not labelled; they occur at the points of the velocity minima.

The relationship between potential and kinetic energy and cross-sectional area is that (my emphasis):

> (1) "**potential energy** density **is proportional to the area** and to the square of the sound pressure" (Mrayati *et al.* 259)
> (2) "**kinetic energy** density **is** proportional to the square of the flow velocity and **inversely proportional to the area**" (Mrayati *et al.* 259)

So, the effect of a constriction (i.e., a reduction in area) is to increase the kinetic energy and to decrease the potential energy. (If this seems counterintuitive, it's because we're talking here not about the effect of a change of volume on a closed cylinder in which a gas is *contained*, which is what we're used to thinking about in introductory chemistry courses, but about the size of an open tube through which a gas is *travelling*.) The effect of producing a constriction differs in different areas of the vocal tract because the initial conditions differ in different areas of the vocal tract, as illustrated in Chiba and Kajiyama (1941).

The equation given in Mrayati *et al.* for calculating the effect of a change in area on a formant frequency is

$$\frac{\Delta F}{F} = \sum_{n=1}^{N} \frac{KE_n - PE_n}{TE_n} \frac{\Delta A_n}{A_n}$$

where

F	=	formant frequency
KE	=	kinetic energy
PE	=	potential energy
TE	=	total energy
A	=	area

Consider the case where kinetic energy is large and potential energy is small: the value of the expression *KE–PE* will be large and positive. If the change in area is a decrease (rather than an increase), then the value of ΔA_n will be negative, the value of the expression $\Delta A_n/A_n$ will be negative, and the change in the frequency of the formant will consist of a large decrease.

Such is the situation at the lips. Each of the resonant frequencies of the vocal tract has a velocity maximum and a pressure minimum at this point. Since potential energy is proportional to (the square of) the sound pressure, and pressure is at a minimum at this point, the potential energy is low. Kinetic energy is proportional to (the square of) flow velocity, and velocity is at a maximum at this point, so the kinetic energy is high. Therefore, the model predicts that the effect of a constriction at the lips should be a lowering of all of the first three resonant frequencies of the vocal tract, as is in fact the case.

As a further example, consider the effect of a constriction in the palatal area. Examining the figures in Chiba & Kajiyama (1941), we see that for the first resonant frequency, velocity is quite high, though not quite as high as it is at the lips. The second resonant frequency has a velocity minimum and a pressure maximum in this area. The situation regarding the first formant will be similar to that described above, except that the effects of the constriction will be just slightly less than at the lips, since velocity at this point is slightly less than at the lips. For the second resonant frequency, since kinetic energy is proportional to velocity and velocity is at a minimum at this point, the kinetic energy is low. Since potential energy is proportional to pressure and pressure is at a maximum at this point, the potential energy is high. Thus, the value of the expression *KE–PE* will be large and negative. If the change in area is a constriction, i.e. a reduction in area, then the term ΔA_n will be negative, the value of the expression $\Delta A_n/A_n$ will be negative, and the effect of the constriction will be a large increase in the frequency of the second formant. The model thus predicts that a constriction in the palatal area should produce (a) an F1 which is quite low, but not quite as low as that produced by a constriction at the lips, and (b) a high F2. Both of these predictions are correct.

References

Abry, Christian; Louis-Jean Boe; and Jean-Luc Schwartz. (1989) Plateaus, catastrophes and the structuring of vowel systems. Journal of phonetics (17)47-54.

Badin, Pascal Perrier; Louis-Jean Boë; and Christian Abry. (1990) Vocalic nomograms: acoustic and articulatory considerations upon formant convergences. Journal of the Acoustical Society of America 87(3)1290-1300.

Beckman, Mary E.; Tzyy-Ping Jung; Sook-hyang Lee; Kenneth DeJong; Ashok K. Krishnamurthy; Stanley C. Ahalt; K. Bretonnel Cohen; and Michael J. Collins. (1995) Variability in the production of quantal vowels revisited. Journal of the Acoustic Society of America 97(1)471-490.

Blankenship, Barbara; Peter Ladefoged; Peri Bhaskararao; and Nichumeno Chase. (1993). Phonetic structures of Khonoma Angami. UCLA working papers in phonetics (84)127-142.

Bradlow, Ann R. (1993). Language-specific and universal aspects of vowel production and perception: a cross-linguistic study of vowel inventories. Ithaca, NY.: Cornell University.

Carré, Rene and M. Mrayati. (1990) Articulatory-acoustic-phonetic relations and modeling, regions and modes. Speech production and speech modelling. Hardcastle, William J. and Alain Marchal, eds. Netherlands: Kluwer Academic Publishers.

Chiba, T., and M. Kajiyama. (1941) The vowel: its nature and structure. Tokyo: Kaiseikan.

Fant, Gunnar. (1960) Acoustic theory of speech production. Netherlands: Mouton.

Hoffman, Paul R.; Gordon H. Schuckers; and Raymond G. Daniloff. (1989) Children's phonetic disorders: theory and treatment. Boston: College-Hill Press.

Honda, Kiyoshi. (1994) Somatoneural relation in the auditory-articulatory linkage. ATR Human Information Processing Research Laboratories manuscript.

Honda, Kiyoshi.; H. Hirai; and N. Kusakawa. (1993) Modeling vocal tract organs based on MRI and EMG observations and its implication on brain function. Annual bulletin of the Research Institute in Logopedics and Phoniatrics (27)37-49. Tokyo.

Jackson, Michel T.T. (1988) Phonetic theory and cross-linguistic variation in vowel articulation. UCLA working papers in phonetics 71.

Jacob, Stanley W. and Clarice A. Francone. (1965) Structure and function in man. Philadelphia: W.B. Saunders.

Johnson, Keith A. (1994) Acoustic phonetics for linguists. Ohio State University manuscript.

Kakita, Yuki; Osamu Fujimura; and Kiyoshi Honda. (1985). Computation of mapping from muscular contraction patterns to formant patterns in vowel space. Phonetic linguistics: essays in honor of Peter Ladefoged, pp. 133-144. Victoria Fromkin, ed. Academic Press.

Kent, Raymond D.; Bishnu S. Atal; and Joanne L. Miller. (1991) (Commentary on) vocal tract and acoustic relationships. Papers in speech communication: speech production, pp. 267-268. R.D. Kent *et al.*, eds. Woodbury, New York: Acoustical Society of America.

Ladefoged, Peter. (1993). A course in phonetics. Harcourt Brace Jovanovitch.

Ladefoged, Peter, and Anthony Bladon. (1982) Attempts by human speakers to reproduce Fant's nomograms. Speech communication (1)185-198.

Ladefoged, Peter, and Mona Lindau. (1989) Modeling articulatory-acoustic relations: a comment on Stevens' "On the quantal nature of speech." Journal of Phonetics (17)99-106.

Laitman, Jeffrey T. (1984) The anatomy of human speech. Natural history 8/84:22-26.

Laitman, Jeffrey T. and J.S. Reidenberg. (1988) Advances in understanding the relationship between the skull base and larynx with comments on the origin of speech. Human evolution 3(1-2)99-109.

Lieberman, Philip. (1975) On the origins of language. The Macmillan series in physical anthropology. New York: Macmillan.

------. (1984) The biology and evolution of language. Cambridge: Harvard University Press.

------. (1991) Uniquely human: the evolution of speech, thought, and selfless behavior. Cambridge: Harvard University Press.

Lindblom, Bjorn, and J. Sundberg. (1971) Acoustical consequences of lip, tongue, jaw, and larynx movements. JASA (50)1166-1179.

Maeda, Shinji. (1990) Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. Speech production and speech modelling. Hardcastle, William J. and Alain Marchal, eds. Netherlands: Kluwer Academic Publishers.

McMinn, R.M.H., and R.T. Hutchings. (1977) Color atlas of human anatomy. Chicago: Year Book Medical Publishers, Inc.

Moore, Christopher A. (1992) The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images. Journal of speech and hearing research 35(5)1009-1023.

Mrayati, M.; R. Carré; and B. Guérin. (1988). Distinctive regions and modes: a new theory of speech production. Speech Communication (7)257-286.

Negulesco, John A. (1993) The anatomy of the head and neck. Department of Anatomy, College of Medicine, Ohio State University.

Perkell, Joseph S. (1969) Physiology of speech production: results and implications of a quantitative cineradiographic study. Cambridge: MIT Press.

Shalev, Michael; Peter Ladefoged; and Peri Bhaskararao. (1993). Phonetics of Toda. UCLA working papers in phonetics (84):89-126.

Smith, T. S. (1971) A phonetic study of the function of the extrinsic tongue muscles. UCLA Working papers in phonetics (18). Los Angeles.

Stevens, Kenneth N. (1972) Quantal nature of speech. Human communication: a unified view, pp. 51-66. E.E. David, Jr. and P.B. Denes, eds. New York: McGraw-Hill.

------. (1989) On the quantal nature of speech. Journal of phonetics (17)3-45.

Stevens, Kenneth N. and Arthur S. House. (1955) Development of a quantitative description of vowel articulation. JASA (27)484-493.

Stone, Maureen; Alice Faber; Lawrence J. Raphael; and Thomas H. Shawker. (1992) Cross-sectional tongue shape and linguopalatal contact patterns in [s], [ʃ], and [l]. Journal of Phonetics (20)253-270.

Swindler, D. R. and C. D. Wood. (1973). An atlas of primate gross anatomy. Seattle: University of Washington Press.

Wood, Sidney. (1982) X-ray and model studies of vowel articulation. Lund University working papers (23).

------. (1986) The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels. JASA 80(2)391-401.